DOI:10.13546/j.cnki.tjyjc.2009.02.011

两个正态总体均值之差的区间估计解析

邱 瑾

（浙江财经学院 数学与统计学院，杭州 310018）

摘 要：文章对两个正态总体均值之差的区间估计进行教学上的解析，尤其针对非均衡样本下方差未知且不相等情形，利用例举对比法说明应如何选择合适的枢轴量。

关键词：两个正态总体；非均衡样本；区间估计

中图分类号：O212 文献标识码：A 文章编号：1002-6487(2009)02-0154-02

1 问题的提出

很多学生在学习两个正态总体均值之差的区间估计时，感觉非常复杂，难于掌握其中的本质。笔者结合在数理统计教学过程中积累的经验对此问题进行解析。

设两个正态总体：$X\sim N(\mu_1,\sigma_1^2)$，$X\sim N(\mu_2,\sigma_2^2)$。$(X_1,X_2,\cdots,X_{n_1})$为来自于总体 X 的随机样本，$(Y_1,Y_2,\cdots,Y_{n_2})$为来自于总体 Y 的随机样本，且两个随机样本相互独立。记

$$\bar{X}=\frac{1}{n_1}\sum_{i=1}^{n_1}X_i,S_1^2=\frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i-\bar{X})^2$$

$$\bar{Y}=\frac{1}{n_2}\sum_{i=1}^{n_2}Y_i,S_2^2=\frac{1}{n_2-1}\sum_{i=1}^{n_2}(Y_i-\bar{Y})^2$$

考虑两个正态总体均值之差，即 $\mu_1-\mu_2$ 的区间估计时，首先分方差已知和未知两种情形。

2 方差已知

方差已知情形非常简单，以样本均值之差的标准化随机变量作为枢轴量，利用该枢轴量服从标准正态分布即可推导出总体均值之差的区间估计。即构造枢轴量

$$U=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\sim N(0,1) \quad (1)$$

可得 $\mu_1-\mu_2$ 的 $1-\alpha$ 置信区间。

3 方差未知

方差未知情形较复杂，具体又细分为三种情况：(1)$\sigma_1^2=\sigma_2^2=\sigma^2$ 未知；(2)$\sigma_1^2\neq\sigma_2^2$，但 $n_1=n_2$；(3)$\sigma_1^2\neq\sigma_2^2$，且 $n_1\neq n_2$。下面对 3 种情况分别进行解析。

3.1 $\sigma_1^2=\sigma_2^2=\sigma^2$ 未知

对于此种情况，主要思想是利用两个样本的合并方差作为 σ^2 的估计，构造 t 统计量做区间估计，即构造枢轴量

$$T_1=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_{pooled}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t(n_1+n_2-2)$$

其中

$$S_{pooled}=\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$$

3.2 $\sigma_1^2\neq\sigma_2^2$，但 $n_1=n_2=n$

由于是均衡样本，因此类似于成对数据的比较，可先做配对差，即令 $D_i=X_i-Y_i$，i=1,2,…,n。且因为 $D_1,D_2,\cdots,D_n$ 是相互独立的，所以可将其视为来自于正态总体的随机样本。注意到此时已经归属于单个正态总体 $N(\mu_1-\mu_2,\ \sigma_1^2+\sigma_2^2)$方差未知的情形，因此用$\bar{D},S_D^2$分别作为 $\mu_1-\mu_2$ 和 $\sigma_1^2+\sigma_2^2$的估计量，构造枢轴量。

$$T_2=\frac{\bar{D}-(\mu_1-\mu_2)}{S_D/\sqrt{n}}\sim t(n-1)$$

其中

$$\bar{D}=\frac{1}{n}\sum_{i=1}^{n}D_i=\bar{X}-\bar{Y},S_D^2=\frac{1}{n-1}\sum_{i=1}^{n}(D_i-\bar{D})^2$$

3.3 $\sigma_1^2\neq\sigma_2^2$，且 $n_1\neq n_2$

若 n_1 和 n_2 均较大，将(1)式中的 σ_1^2 与 σ_2^2 分别用其估计量 S_1^2 与 S_2^2 替代，则利用中心极限定理和 Slutsky 定理，构造枢轴量。

$$U^*=\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}\xrightarrow{d}N(0,1)$$

若 n_1 和 n_2 均较小，Welsh(1949)得到 U* 近似服从 t(m) 分布，其中

基金项目：国家社会科学基金项目(07CTJ001)；浙江省哲学社会科学规划常规性课题(06CGYJ21YQB)

$$m=\frac{\tilde{S}^4}{\frac{S_1^4}{n_1^2(n_1-1)}+\frac{S_2^4}{n_2^2(n_2-1)}},\tilde{S}^2=\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}$$

若 m 不为整数时，利用线性插值法可求出临界值，或者用最接近 m 的整数即可。

当 n_1 和 n_2 均较小时，参考文献[2]中的主要思想是构造配对，更易让学生接受。但是由于是非均衡样本，如何构造才是合理可行，且充分利用样本信息是我们考虑的重点。对比以下几种学生容易想到的统计量我们来进行分析。不妨设 $n_1<n_2$。

(1)令 $\xi_i=X_i-Y_i, i=1,2,\cdots,n_1$。这是沿用了配对思想，可直接利用情况 3.2 的结果。但是我们损失了 $n_2<n_1$ 个样本观测信息$(Y_{n_1+1},\cdots,Y_{n_2})$。

(2)令 $\eta_i=X_i-\bar{Y}, i=1,2,\cdots,n_1$。此时虽然

$$E\eta_i=\mu_1-\mu_2, Var\eta_i=\sigma_1^2+\frac{1}{n_2}\sigma_2^2, i=1,2,\cdots,n_1$$

但是

$$Cov(\eta_i,\eta_j)=\frac{1}{n_2}\sigma_2^2\neq 0, i\neq j$$

因此，不能将 $\eta_1,\eta_2,\cdots,\eta_n$ 看作来自于正态总体 $N(\mu_1-\mu_2,\sigma_1^2+(1/n_2)\sigma_2^2)$的随机样本，则以样本均值$\bar{\eta}$作为 $\mu_1-\mu_2$ 的估计来利用单个正态总体的结果不再可行。

(3)参考文献[2]中给出了如下配对：令

$$\gamma_i=X_i=\sqrt{\frac{n_1}{n_2}}Y_i+\frac{1}{\sqrt{n_1n_2}}\sum_{i=1}^{n_1}Y_i-\bar{Y}, i=1,2,\cdots,n_1,$$ 则

$$E\gamma_i=\mu_1-\mu_2, Var\gamma_i=\sigma_1^2+\frac{n_1}{n_2}\sigma_2^2, i=1,2,\cdots,n_1,$$

更重要的是

$$Cov(\gamma_i,\gamma_j)=0, i\neq j$$

因此，可以将 $\gamma_1,\gamma_2,\cdots,\gamma_n$ 看作来自于正态总体 $N(\mu_1-\mu_2,\sigma_1^2+(n_1/n_2)\sigma_2^2)$的随机样本。令

$$\bar{\gamma}=\frac{1}{n_1}\sum_{i=1}^{n_1}\gamma_i=\bar{X}-\bar{Y}, S_\gamma^2=\frac{1}{n_1-1}\sum_{i=1}^{n_1}(\gamma_i-\bar{\gamma})^2$$

注意到

$$E\bar{\gamma}=\mu_1-\mu_2, Var\bar{\gamma}=\frac{1}{n_1}\sigma_1^2+\frac{1}{n_2}\sigma_2^2$$

则以$\bar{\gamma}$作为 $\mu_1-\mu_2$ 的估计，S_γ^2作为 $Var\bar{\gamma}$的估计，再利用单个正态总体的结果可得

$$T_3=\frac{\bar{\gamma}-(\mu_1-\mu_2)}{S_\gamma/n_1}\sim t(n_1-1)$$

从而可得 $\mu_1-\mu_2$ 的 $1-\alpha$ 置信区间。

在以上各种情形的教学中始终强调枢轴量的结构，而避免让学生去记忆整理后的表达式。同时强调用配对方法时，本质上是在寻找独立同分布随机变量，从而构成单个正态总体的一个随机样本。

参考文献：

[1]Welch, B. Further notes on Mrs. Aspin's tables [J].Biometrika, 1949,36.

[2]孙荣恒.应用数理统计[M].北京：科学出版社,2003.

[3]陈希孺.高等数理统计学[M].合肥：中国科技大学出版社,1999.

[4]茆诗松，王静龙，濮晓龙.高等数理统计[M].北京：高等教育出版社,2000.

[5]茆诗松，王静龙.数理统计[M].上海：华东师范大学出版社,1990.

（责任编辑/浩 天）